Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1249 - 1257
Received 20 January 2021; Accepted 08 February 2021.

A Novel Approach to Classify Train Reviews Based on Sentiment Analysis and Compare the Probability of Error Rate over Hadoop Architecture

[1]CH. Sai Ravi Teja, [2]S. Stewart Kirubakaran, [3]R.Senthil Kumar

[1]UG Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai

[2,3]Assistant Professor, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai

[1]tejacherukuri111@gmail.com, [2]stewartkirubakaran@gmail.com,[3]rsenthilmecse@gmail.com

ABSTRACT:

Big Data and Sentiment Analysis are drawn big attention in public social data due to complex analysis of large volume of data. This paper presents here sentiment analysis and classification of train data using Hadoop Architecture. Trains are the transportation mode for 40 per cent of the world 's passengers. Trains are unlike buses that get into frequent traffic jams, but even trains can be delayed due to problems with engines, engineering problems, problems with passengers or natural disasters. The passengers face different problems due to train delays.While we cannot avoid any delays, we can forecast the delays by observing the conditions of the train stations, such as weathers. There are also other issues such as how passengers will be treated on the waiting list. Such data sets would render a large volume of data waiting to be analyzed without data incoherence.The Big Knowledge Technology uses all of the information to approximate once and to predict how the delay can occur, so that some could use different trains to travel. In addition, we will use time ticketing schemes on the train ticketing schemes so that the passenger roster will be managed in our future proposal by Spark technology.

Keywords: Transportation, passengers,natural disasters,approximate,ticketing schemes.

INTRODUCTION:

Data beyond the ability of storage and beyond the power of processing these data is called Big Data. Big data really means big data; it is a series of massive databases which can't be processed using conventional computing techniques. Big data is not just a data but a complete subject,that involves various methods, techniques and frameworks. Big Data is called data which is very large in scale. We are usually operate the size and data MB(Wordbook, Excel) or maximum GB(Movies, Codes), but data in Petabyte’s is called Big Data, i.e. 10 ^ 15 byte sizes. It is stated that nearly 90% of today 's data were generated in the last 7 years’ data should be presented in order to transform the process to words the past development.

SCOPE OF THE PROJECT:

Within this paper we analyze data from the Railway Network using a Hadoop method along with some Hadoop ecosystems including hdfs, MapReduce, sqoop, hive and pig. By using these present tools, data processing is possible without any constraints, no data lost problem, we can achieve a high throughput,Even very low maintenance costs because it's an open source program, it's compatible with all platforms because it's Java-based. In Railway Network info, large volume of research paper publishing website storage is linked.

LITERATURE REVIEW:

1)Multiport Railway Power Conditioner and its Clean Energy Access Management Strategy [1]

A multi-port railway power conditioner (MP-RPC) with access to renewable energy is proposed to achieve the mutual benefits between railway systems and renewable energy systems (RESs) in order to realize on-site power usage and compensation for negative series. Its power-flow management strategy is proposed and analyzed for MP-RPC with 5 classic operating modes.P-RPC's main component is tri-port isolated DC / DC Converters which can be used for insulation and transmission of electricity.

The port power control strategy with feed forward decoupling is implemented by small-signal modeling and study of the three-port converter.Based on the feed forward decoupling matrix, the coupling mechanism is decoupled into two independent control loops, which can effectively boost mechanism dynamic performance. Lastly, the traction power platform and virtual prototype are installed in the Lab and MP-RPC's topology and control methods are effectively tested.

2)Prediction of device failure based on a Markov Bayesian model of the network [2]

Because of the complexity of software products and production processes, software reliability models need to provide several parameter handling capacity. They should also provide flexibility in model construction in terms of information updating to adapt to the constantly refreshed data.Under this sense the current models of software reliability are not versatile. The key explanation for that is that the models are correlated with several static assumptions.

Received 20 January 2021; Accepted 08 February 2021.

Bayesian network is a valuable tool for solving this problem since it possesses a good ability to adapt to problems involving complex variant factors.In this paper, a machine prediction model is built based on Markov Bayesian networks, and a method is proposed to solve the network problem.

3) Inductive thinking when diagnosing faults in network [3]

There is a need for autonomous systems that must be able to work with minimal human interference to detect and isolate faults, and to recover from these faults. Within this paper we present a new hybrid model-based architecture and data clustering (MDC) for fault monitoring and diagnostics.Suitable for dynamic complex systems with continuous and discrete variables. MDC approach allows both structure and parameters of defined models to be modified using supervised and reinforced learning techniques. The MDC solution will be demonstrated using the model and data from the NASA Ames Research Center's Hybrid Combustion Facility (HCF).

MODULES

- Existing Application (MySQL)
- Connector (Sqoop)
- Analysis Query Language (Hive)
- Analysis Latin Script (Pig)
- Processing (MapReduce)

MODULE DIAGRAMS & MODULE DESCRIPTION

1)Existing Application (MySQL):

The MySQL is a method of relational database management. RDBMS uses relationships or tables to store data from the Railway Network as a column matrix of primary key rows. With MySQL, Railway Network data can be obtained, stored, analyzed, retrieved, extracted and used mainly for commercial purposes.Current concept deals with providing backend using MySQL which involves a lot of drawbacks i.e. data drawback is that processing time is high when the data is enormous and once data is lost we cannot recover so we propose concept using Hadoop device.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1249 - 1257
Received 20 January 2021; Accepted 08 February 2021.

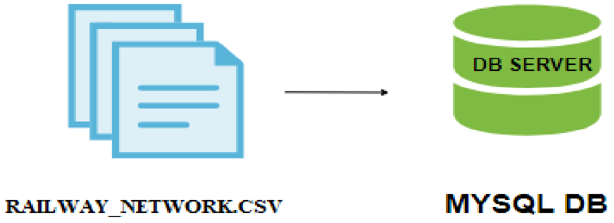

Figure1: It shows that railway network.csv file transfer the data through MySQL DB.

2) **Connector (Sqoop):**

Sqoop is a command line interface framework for sharing data between relational databases (MySQL) and Hadoop on the Railway Network. Here in MySQL database, the Railway Network data must be imported via Sqoop to HDFS. Railway network data can be transferred from MySQL to HDFS / Hive, and then the java classes are created.In previous cases, data flow had been from RDBMs to HDFS. We can import data from HDFS into RDBMs using the "Export" tool. Sqoop fetches table metadata from MySQL database before exporting. Therefore, we need to create a table with the metadata needed first.

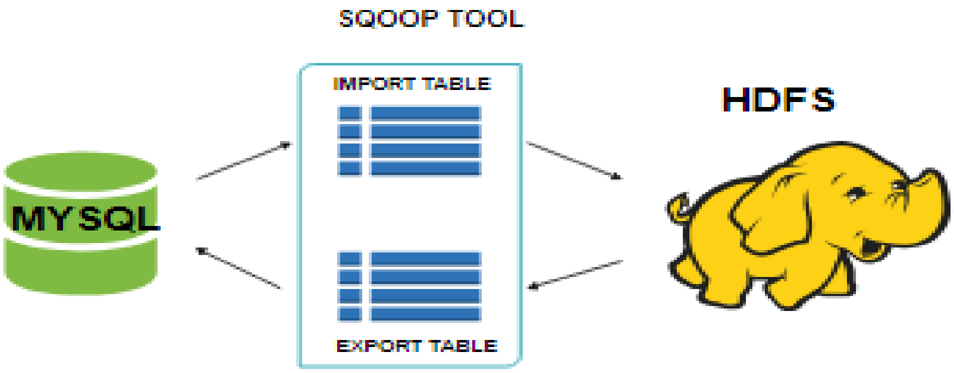

Figure 2: Data transformation between the MySQL and HDFS.

3) **Analysis Query Language (Hive)**

Hive is a Hadoop data warehouse framework that runs SQL like HQL (Hive query language) queries that are transformed internally to map reduces work. Within Hive, data tables and databases for the Railway Network are first generated and then loaded into these tables. Data warehouse Hive as Railway Network designed for handling and querying only structured data stored in tables.Hive arranges data tables for the Railway Network into partitions. This is a way to segment a table into related sections, based on partitioned column values. Using partition, a portion of the given dataset is easily queried. Tables or partitions are subdivided into seals to

Received 20 January 2021; Accepted 08 February 2021.

give the Railway Network extra data structure that can be used for more effective queries.Bucketing works based on the hash value function of any table column.

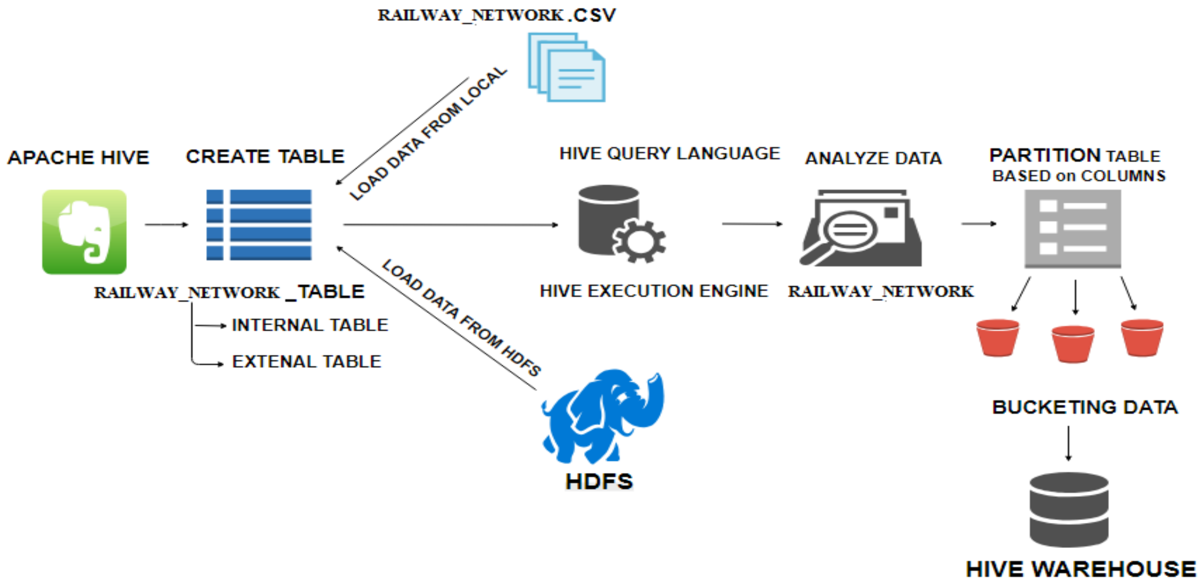

Figure 3: Way of transformation between each data process should be separated equally.

4) **Processing (MapReduce)**

MapReduce is a system that allows us to write applications for the efficient processing of massive quantities of Railway Network data on large commodity hardware clusters in parallel. MapReduce is a processing technique and program model for the Java-based distributed computing.The MapReduce algorithm includes two important tasks: Map and Reduce. MapReduce program executes in three stages, namely stage mapping, stage shuffling, and stage reduction. The role of the map or mapper is to process data from the inputs. The input data is usually in the form of a file or directory, and is stored in the Hadoop file system (HDFS). The input file is passed line by line into the mapper function.The mapper processes the data, generating a few tiny pieces of data. This step is the combination of step Shuffle and stage Reduce. The job of the Reducer is to process the mapper-sourced data. It generates a new collection of output after processing, which will be stored in the HDFS**.**

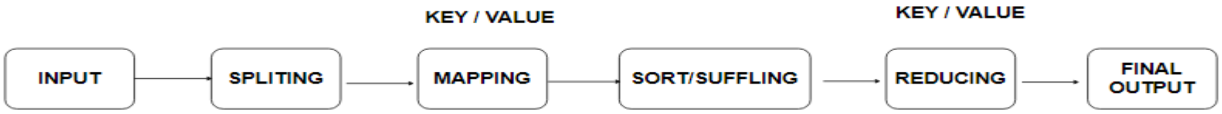

Figure 4: flow process of the mapping

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1249 - 1257
Received 20 January 2021; Accepted 08 February 2021.

PROPOSED SYSTEM:

The current definition relates to the provision of We can use Hadoop tool to database Analyze no data limitations and clearly Add machine number to the cluster and We are having results with less time, high Durability and repair costs are much lower So we use the joins, groups so Bucketing on Hadoop techniques.

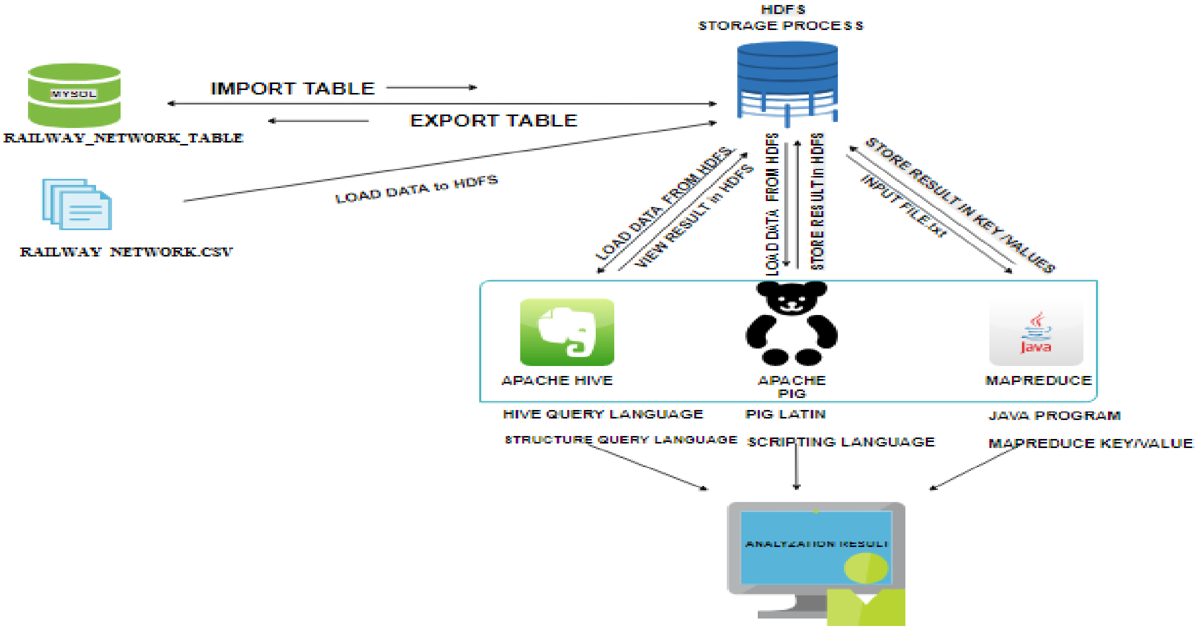

Figure 5: System Architecture.

RESULT AND DISCUSSION

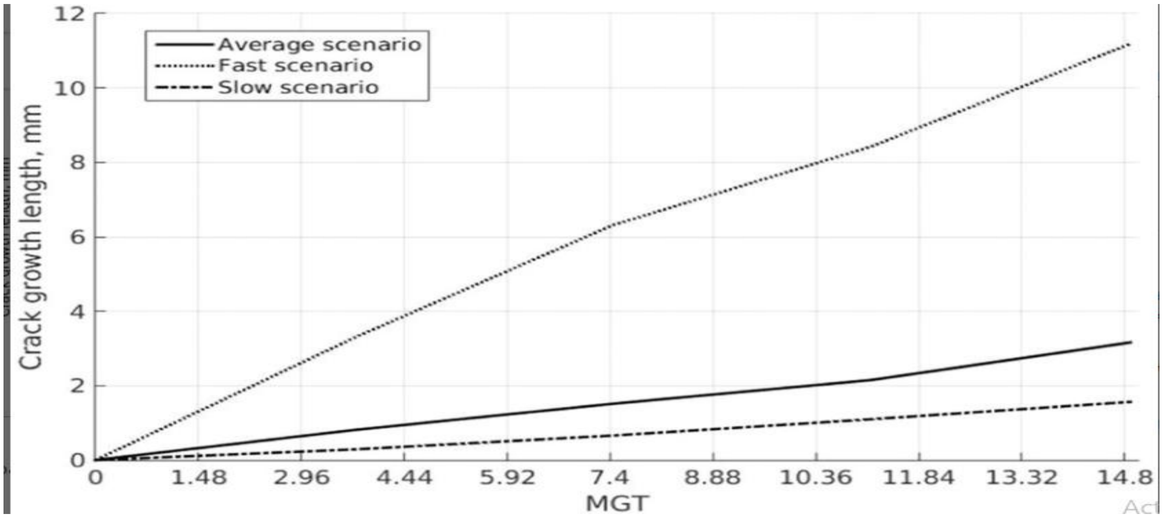

Figure 6: It shows the relation between the average scenario, fast scenario and slow scenario.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1249 - 1257
Received 20 January 2021; Accepted 08 February 2021.

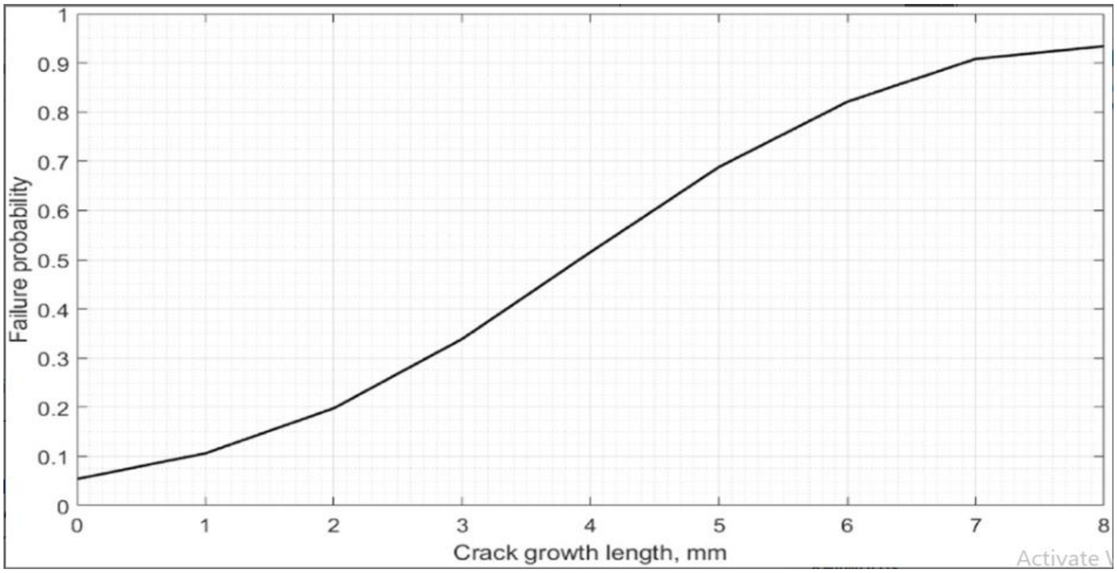

Figure 7: Probability of rail failure based on the growth of crack length.

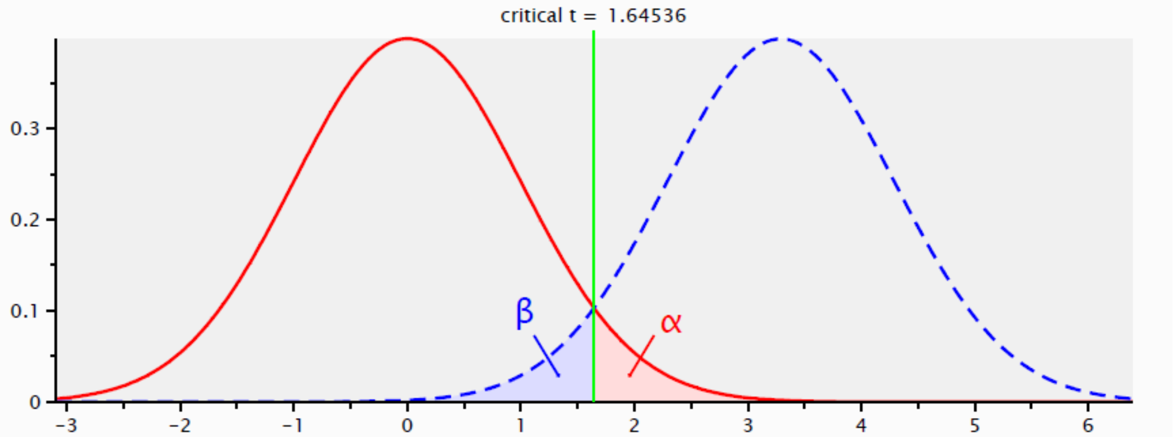

Figure8: Probability of rail failure based on the critical value

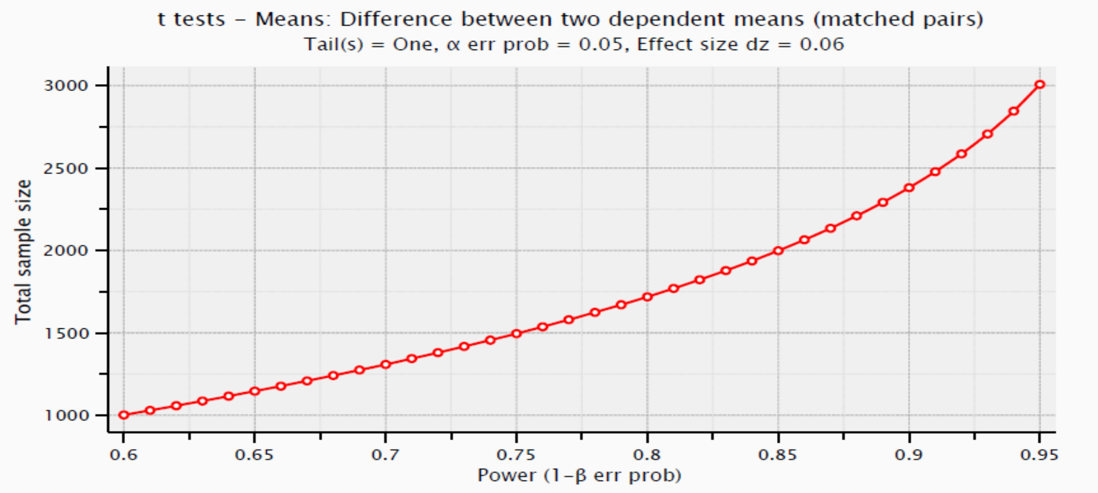

Figure9: Mean and Standard Deviation for the Sample Sizes

Received 20 January 2021; Accepted 08 February 2021.

CONCLUSION:

Data classification is very difficult to determine the sentiment of user review or tweets whether it is positive, negative or neutral case. Using big data with Hadoop architecture the data is classified for Railway Network to enhance customer experience.In this paper, we presented data and analysis on the Railway Network within relation to the research article on railway networks. Analyze the data from the Railway Network in the Hadoop ecosystem and boost travel based on accident prediction and alternate routes. Hadoop ecosystem has hive, pig, map reduces resources to process if production takes less time to processand outcome is going to be very fast. Thus, in this project, data from the Railway Network, which will traditionally be stored in RDBMS, will be reduced in performance, hence, by using the Hadoop tool, data will be processed more quickly and efficiently.

FUTURE ENHANCEMENT:

Apache Spark is an open source processing engine composed of speed, use case, and analytics. If you have large quantities of data that cannot be supported by a standard Map Reduce system requiring low latency processing, then Spark is the alternative. Spark provides fast lightning speed in-memory cluster computing, and supports Java, Scala, and Python APIs for easy development**.**

REFERENCES:

[1] H. Hu, Z. He, X. Li, K. Wang, and S. Gao, “Power-quality impact assessment for high-speed railway associated with high-speed trains using train timetable—Part I: Methodology and modeling.

[2] G. Qiao, N. Ding, S. Zhou, and K. Yu, “Power quality conditioner for high-speed railway based on traction transformer with V/v wiring.

[3] J. W. Kolar and G. Ortiz, “Solid-state-transformers: Key components of future traction and smart grid systems.

[4]Z. Zhang, B. Wu, J. Kang, and L. Luo, “A multi-purpose balanced transformer for railway traction applications.

[5] H. Akagi and R. Kitada, “Control and design of a modular multilevel cascade BTB system using bidirectional isolated dc/dc converters.

[6] H. Tao, J. L. Duarte, and M. A. M. Hendrix, “Three-port triple-half-bridge bidirectional converter with zero-voltage switching.

[7] J. E. Huber and J. W. Kolar, “Solid-state transformers: On the origins and evolution of key concepts.

[8] L. Costa, G. Buticchi, and M. Liserre, “Quad-active-bridgedc-dc converter as cross-link for medium voltage modular inverters.

[9] L.X. Wang, "Fuzzy systems are universal approximators:' Proceedings of IEEE International Conference on Fuzzy Systems.

[10] D. Imamori and K. Tajima, “Predicting popularity of twitter accounts through the discovery of link-propagating early adopters.